COMPARISONS OF THEIL'S AND SIMPLE REGRESSION ON NORMAL AND NON-NORMAL DATA SET WITH DIFFERENT SAMPLE SIZES

[1]ESEMOKUMO PEREWAREBO AKPOS, [2]OPARA, JUDE

[1]Department of Statistics, School of Applied Science, Federal Polytechnic Ekewe, Yenagoa, Bayelsa State, Nigeria
[2]Department of Statistics, Imo State University, PMB 2000, Owerri Nigeria
E-mail: [1]contactperes4good@gmail.com, [2] judend88@yahoo.com

Abstract - This paper is on comparisons of Theil's and simple regression on normal and non-normal data set with different sample sizes. Data used for this study were collected from a real life practical conducted by the researchers in their homes on the weight of soap and the number of days it had been used. Thus dependent variable(y) is weight (grams) of the soap and independent variable is the number of days (x). To know the efficiency of one method over the other, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mean Square Error (MSE) were used. From the analysis, the result revealed that there is a significant relationship between dependent and independent variables for both the parametric OLS regression and non-parametric Theil's regression with and without residual normality validity. Hence, an inverse relationship between x and y, that is as the number of days increase, weight of the soap decreases. It can be concluded that the parametric OLS regression performs better than its non-parametric Theil's regression since their Residual standard error, AIC and BIC values are all smaller for both the normal and non-normal real data. The result of the real life data was used for data simulation of sample sizes of n = 30, 50, 100, 150, 200, 400, 500, 700, 900, 1000, and 1500, and the results revealed that the parametric OLS regression performs better than its non-parametric Theil's regression since their EMS, AIC and BIC values are all smaller. It can be concluded that the regression line gave a good fit to the observed data since the line explains over 99% of the total variation of the Y values around their mean for both models. Even though the both models are good in this study, the OLS is more efficient. Therefore the researchers recommend that future research should look at a similar work with both high and low coefficient of variation of different sample sizes with normal and non-normal data, and also with more than one explanatory variable to examine the differences between the parametric and nonparametric Regression.

Keywords - Theil's Regression, Simple Regression, Anderson-Darling technique, Akaike Information Criterion, Bayesian Information Criterion, Error Mean Squares

I. INTRODUCTION

Regression analysis is a statistical technique that express mathematically the relationship between two or more quantitative variables such that one variable (the dependent variable) can be predicted from the other or others (independent variables). Regression analysis is very useful in predicting or forecasting (Inyama and Iheagwam, 2006). It can also be used to examine the effects that some variables exert on others. However, regression analysis may be simple linear, multiple linear or non linear. In this study, simple linear case is applicable with its non-parametric equivalent. The simple linear regression model is the ordinary or traditional equation representing the relationship between two variables; the response and the explanatory variables. Sometimes the residuals in a regression analysis may deviate far from the others. In this case, an outlier occurs. It is obvious that no observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. Therefore, the probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions. Observations that in the opinion of the investigator stand apart from the bulk of the data have been called "outliers", "extreme observations" "discordant observations", "rouge values", "contaminants", "surprising values", "mavericks" or "dirty data" (Ranjit; 2005). An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs. An outlier is a data point that is located far from the rest of the data.

Again, the presence of outliers may contribute to non-normal distribution. Consider a situation where the distribution of the errors is not normal. If the errors are coming from a population that has a mean of zero, then the OLS estimates may not be optimal, but they at least have the property of being unbiased. If we further assume that the variance of the error population is finite, then the OLS estimates have the property of being consistent and asymptotically normal. However, under these conditions, the OLS estimates and tests may lose much of their efficiency and they can result in poor performance (Mutan; 2004). To deal with these situations, two approaches can be applied. One is to try to correct non-normality, if non-normality is determined and the other is to use alternative regression methods, which do not depend on the assumption of the normality (Birkes and Dodge; 1993). In straight-line regression, the least squares estimator of the slope is sensitive to outliers and the associated confidence interval is affected by non-normality of the dependent variable. A simple and robust alternative to least squares regression is Theil regression, first proposed by Theil (1950).

Theil's method actually yields an estimate of the slope of the regression line.

II. REVIEW OF RELATED LITERATURES

There is need to review works done by past researchers in order to have a proper guide. Here are some recent works done by past researchers.

Opara et al (2016) researched on the comparison of parametric and non-parametric linear regression. First, the set of data was subjected to normality test, and it was concluded that all errors in the y-direction are normally distributed (i.e. they follow a Gaussian distribution) for the commonly used least squares regression method for fitting an equation into a set of (x,y)-data points using the Anderson-Darling technique. Data used for the study were collected from a trader in Dauglas Owerri Market in Imo State Nigeria who sales pears. The numbers of rotten pears (y) in 20 randomly selected boxes from a large consignment were counted after they have kept in storage for a studied number of days (x). The use of a programming language software known as "R Development" and Minitab were used in the study. From their analysis, the result revealed that there exists a significant relationship between the numbers of rotten pears and the number of days for both the ordinary least squares and the Theil's regression. It was concluded that the parametric OLS is better than its non-parametric Theil's regression since their AIC and BIC are both lower than that of Theil's regression.

Okenwe et al (2016) worked on Parametric Versus Non-Parametric Simple Linear Regression on Data with and Without Outliers. Data used for the study were collected from the department of Mass Communication, Imo State University Owerri Imo State Nigeria. Twenty five (25) students were selected at random to determine the Cumulative Grade Point Average (CGPA) at the end of 2014/2015 Academic session (Y) and their respective Joint Admission Matriculation Board (JAMB) score (X). The set of data was subjected to normality test, and it was concluded that all residuals in the y-direction are not normally distributed via the Anderson-Darling technique. The data after removing outliers were re-analyzed. From the analysis, the result revealed that there was a significant relationship between students CGPA and their JAMB scores for both the parametric OLS regression and non-parametric Theil's regression with and without outliers. It was concluded that the parametric OLS is better than its non-parametric Theil's regression for both data with and without outliers since their standard error, AIC and BIC are lower than that of Theil's regression. It was also concluded that the standard error for the parametric regression with outliers which is 0.3405 reduced to 0.1962 for the parametric regression without outliers. On the other hand, the standard error for the non-parametric regression with outliers which is 0.3609 reduced to 0.2087 for the non-parametric regression without outliers. It implied that the model for the data without outliers is more efficient than the model for the data with outliers for both the parametric and non-parametric regression.

Ohlson and Kim (2014) in their work titled "Linear Valuation without OLS: The Theil-Sen Estimation Approach" said that OLS confronts two well-known problems in many archival accounting research settings. First, the presence of outliers tends to influence estimates excessively. Second, in the cross-sections, models often build in heteroscedasticity which suggests the need for scaling of all variables. Their study compared the relative efficacy of Theil (1950) and Sen (1968) (TS) estimation approach vs. OLS estimation in cross-sectional valuation settings. Next-year earnings or, alternatively, current market value determines the dependent variable. To assess the two methods' estimation performance the analysis relied on two criteria. The first focused on the inter-temporal stability of coefficient estimates. The second focused on the methods' goodness-of-fit, that is, the extent to which a particular model's projected values come close to actual values. On both criteria, results showed that TS performed much better than OLS. The dominance was most apparent when OLS estimates have the "wrong" sign. TS estimations, by contrast, never lead to such outcomes. Conclusions remained intact even when variables have been scaled for size.

Ekezie and Opara (2014) researched on Estimation of Bivariate Regression Data via Theil's algorithm. The method was adopted since all errors in the y-direction are not normally distributed (i.e. the do not follow a Gaussian distribution) for the commonly used least squares regression method for fitting an equation into a set of (x,y)-data points using the Kolmogorov Smirnov test. The algorithms for Theils were stated in the study. The data used for their research were collected from selected primary schools in Owerri Municipal, Imo State Nigeria. The data were on weights and shoulder heights of 100 randomly selected pupils in primary four, five and six. The use of a programming language software known as "R Development" was used to write an appropriate expression in the study. From the analysis, the result revealed that there exist a significant relationship between weights and shoulder heights of primary school pupils, and the estimated fitted Theil's is $\hat{y}_i = 42.5833 + 0.1177\ x_i$ and it was observed that both the intercept and slope were significant.

Having reviewed some of these past researches, we shall embark on Parametric Versus Non-Parametric Simple Linear Regression on Data With and without Outliers using real life data of CGPA and JAMB scores.

III. METHODOLOGY

Simple Linear Regression

This is a regression line that involves only two variables as it is applicable in this research study. A widely used procedure for obtaining the regression line of y on x is the Least Squares Method.

The linear regression line or y on x is

$$y = \alpha + \beta x + e \quad \ldots \quad (1)$$

where y is the response or dependent variable, x is the predictor or independent variable. α is the intercept, β is the slope, while e is the error term.

Using the least squares method, the parameters are estimated as shown in equations (2) and (3);

$$\hat{\beta} = \frac{n\Sigma x_i y_i - \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad \ldots \quad (2)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \ldots \quad (3)$$

The calculation is usually set out in Analysis of Variance (ANOVA) table as shown in Table 1

Variance	Degree of freedom	Sum of square	Mean square
Regression	1	RSS = $\beta\Sigma xy$	$RMS = RSS$
Error	n – 2	ESS = TSS – RSS	$EMS = {}^{ESS}/_{n-2}$
Total	n – 1	TSS = Σy^2	

Table 1: Regression Table

The test statistic is given by

$$F_{cal} = \frac{RMS}{EMS} \quad \ldots \quad (4)$$

The F_{cal} is now compared with the F-value obtained from the F-table or F-tabulated with 1 and (n – 2) degree of freedom.

Theil's Regression Method

Theil's regression is a nonparametric method which is used as an alternative to robust methods for data sets with outliers. Although the nonparametric procedures perform reasonably well for almost any possible distribution of errors and they lead to robust regression lines, they require a lot of computation. This method is suggested by Theil (1950), and it is proved to be useful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

Sprent (1993) states that for a simple linear regression model to obtain the slope of a line that fits the data points, the set of all slopes of lines joining pairs of data points (x_i, y_i) and (x_j, y_j), $x_j \neq x_i$, for $1 \leq i < j \leq n$ should be calculated by;

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \quad \ldots (5)$$

Thus b^* is the median of all Equation (5)

Hence, in this study, for n observations, we have $\frac{n(n-1)}{2}$ algebraic distinct $b_{ij} = b_{ji}$

But a^* is the median of all $a_i = y_i - b^* x_i$

The mean square error is given in equation ()

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n - k} \quad \ldots \quad (6)$$

Akaike Information Criterion (AIC)

The Akaike's information criterion AIC (Akaike, 1974) is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. Thus, the AIC is defined as;

$$AIC = e^{\frac{2k}{n}} \frac{\sum \hat{u}_i^2}{n} = e^{\frac{2k}{n}} \frac{RSS}{n} \quad \ldots (7)$$

where *k* is the number of regressors (including the intercept) and *n* is the number of observations. For mathematical convenience, Equation (7) is written as;

$$\ln(AIC) = \left(\frac{2k}{n}\right) + \ln\left(\frac{RSS}{n}\right) \quad \ldots (8)$$

where ln (AIC) = natural log of AIC and $\frac{2k}{n}$ = penalty factor.

Bayesian Information Criterion (BIC)

Bayesian Information Criterion BIC (Schwarz, 1978) is a measure of the goodness of fit of an estimated statistical model and can also be used for model selection. It is defined as

$$BIC = n^{\frac{k}{n}} \frac{\sum \hat{u}_i^2}{n} = n^{\frac{k}{n}} \frac{RSS}{n} \quad \ldots (9)$$

Transforming Equation (3) in natural logarithm form, it becomes (See Equation (9));

$$\ln(BIC) = \frac{k}{n}\ln(n) + \ln\left(\frac{RSS}{n}\right) \quad \ldots (10)$$

where $\frac{k}{n}\ln(n)$ is the penalty factor. For model comparison, the model with the lowest AIC and BIC score is preferred.

Data Analysis

Data used for this study were collected from a real life practical conducted by the researchers in their homes of the weight of soap and the number of days it had been used. Thus dependent variable(y) is weight (grams) of the soap and independent variable is the number of days (x). The data generated from the experiment conducted by the two researchers are presented in Table 1 and Table 2.

i	y	x	i	Y	x	i	y	x
1	13.21	1	10	9.31	10	19	5.51	19
2	13.01	2	11	9.04	11	20	5.27	20
3	12.48	3	12	8.62	12	21	4.91	21
4	12.13	4	13	8.38	13	22	4.51	22
5	11.78	5	14	8.02	14	23	4.09	23
6	11.35	6	15	7.46	15	24	3.46	24
7	10.08	7	16	7.12	16	25	3.17	25
8	10.32	8	17	6.49	17	26	2.89	26
9	10.01	9	18	6.08	18	27	2.34	27

Table 1: Weight (y) and Number of days (x)

i	y	x	i	y	x	i	y	x
1	13.21	1	10	10.15	10	19	7.1	19
2	12.84	2	11	9.802	11	20	6.75	20
3	12.52	3	12	9.46	12	21	6.4	21
4	12.17	4	13	9.14	13	22	6.04	22
5	11.83	5	14	8.8	14	23	5.68	23
6	11.5	6	15	8.48	15	24	5.33	24
7	11.15	7	16	8.13	16	25	4.99	25
8	10.82	8	17	7.78	17	26	4.63	26
9	10.48	9	18	7.43	18	27	4.31	27

Table 2: Weight (y) and Number of days (x)

The data set was subjected to normality test using Anderson-Darling Technique via Minitab software, and it can be concluded that the error term in Table 1 is not normally distributed, while the error term of Table 2 is distributed normally. Hence, we go ahead to analyze the parametric OLS with its non-parametric counterpart for the two Tables using R Software.

Having carried out the analysis with normal and non-normal data for both parametric and non-parametric regression, we can conclude that from the result that there is inverse relationship between x and y, that is as the number of days increase, weight of the soap decreases in both the parametric and non-parametric regression models. Again, it can be concluded that the parametric OLS regression performs better than its non-parametric Theil's regression since their Residual standard error, AIC and BIC values are all smaller for both the normal and non-normal data. The coefficient of determination for OLS versus Theil's for the non-normal data are 99.73% and 99.73% respectively, while that of OLS versus Theil's for the normal data are 99.99% and 99.99% respectively. Hence; it can be concluded that the regression line gives a good fit to the observed data since the line explains 99.73% and 99.99% of the total variation of the Y values around their mean for both models.

Let us go further to perform simulation based on the results obtained above on different sample sizes for the both models. The summary result is displayed in Table 3

		Sample sizes										
	n	30	50	100	150	200	400	500	700	900	1000	1500
OLS	$\hat{\alpha}$	13.714	13.713	13.710	13.709	13.706	13.796	13.706	13.706	13.705	13.706	13.707
	$\hat{\beta}$	-0.415	-0.416	-0.416	-0.420	-0.423	-0.421	-0.421	-0.421	-0.420	-0.420	-0.421
	AIC	-124.628	-214.251	-415.243	-609.179	-825.517	-1629.537	-2043.217	-2938.255	-3798.871	-4227.508	-6323.629
	BIC	-120.424	-208.515	-407.427	-600.147	-815.622	-1617.562	-2030.573	-2924.602	-3784.464	-4212.784	-6307.69
	EMS	0.0008	0.0007	0.0009	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Theil's	a^*	13.746	13.736	13.740	13.738	13.730	13.732	13.733	13.733	13.731	13.732	13.733
	b^*	-0.426	-0.417	-0.413	-0.421	-0.425	-0.422	-0.423	-0.423	-0.422	-0.421	-0.423
	AIC	-117.238	-210.089	-404.789	-597.737	-812.114	-1602.92	-2009.74	-2891.352	-3739.32	-4161.472	-6178.759
	BIC	-113.034	-204.353	-396.972	-588.705	-802.219	-1590.946	-1997.096	-2877.699	-3724.913	-4146.748	-6162.819
	EMS	0.0010	0.0008	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 3: Simulation based on different sample sizes

Looking at the simulated result in Table 3, it can be concluded that there is inverse relationship between x and y, that is as the number of days increase, weight of the soap decreases in both the parametric and non-parametric regression models. Again, it can be concluded that the parametric OLS regression performs better than its non-parametric Theil's regression since their EMS, AIC and BIC values are all smaller.

CONCLUSION

From the analysis, the result revealed that there is a significant relationship between dependent and independent variables for both the parametric OLS regression and non-parametric Theil's regression with and without residual normality validity. Hence, an inverse relationship between x and y, that is as the number of days increase, weight of the soap decreases. It can be concluded that the parametric

OLS regression performs better than its non-parametric Theil's regression since their Residual standard error, AIC and BIC values are all smaller for both the normal and non-normal real data. The result of the real life data was used for data simulation of sample sizes of n = 30, 50, 100, 150, 200, 400, 500, 700, 900, 1000, and 1500, and the results revealed that the parametric OLS regression performs better than its non-parametric Theil's regression since their EMS, AIC and BIC values are all smaller. From the result of this study, it can be concluded that the regression line gives a good fit to the observed data since the line explains over 99% of the total variation of the Y values around their mean for both models. Even though the both models are good in this study, the OLS is more efficient. Therefore the researchers recommend that future research should look at a similar work with both high and low coefficient of variation of different sample sizes with normal and non-normal data to examine the differences between the parametric and nonparametric Regression. The researchers further recommend that future researchers should study a similar work by examining a regression model with more than one explanatory variable for both the parametric and non-parametric cases.

REFERENCES

[1] Akaike, H. *(1974),* "A new look at the statistical model identification" *(PDF), IEEE Transactions on Automatic Control 19 (6): 716–723,* doi*:*10.1109/TAC.1974.1100705, MR 042371

[2] Birkes, D., and Dodge, Y.(1993). Alternative Methods of Regression. New York, NY: Wiley.

[3] Dietz, E. J. (1989). Teaching Regression in a Nonparametric Statistics Course. The American Statistician. 43, 35-40.

[4] Ekezie, D. D., and Opara, J. (2015). Estimation of Bivariate Regression Data Via Theil's Algorithm. Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 5(8): 29-34© Scholarlink Research Institute Journals, 2014 (ISSN: 2141-7016).

[5] Inyama, S.C. and Iheagwam, V.A. (2006): Statistics and Probability. A Focuse on Hypotheses Testing. Third edition. Strokes Global Ventures Owerri, Imo State, Nigeria.

[6] Mutan, O.M. (2004). Comparison of Regression techniques via monte carlo simulation. A thesis submitted to the school of natural and applied sciences of middle east technical University.

[7] Ohlson, J.A., and Kim, S. (2014). Linear valuation without OLS: The Theil-Sen Estimation Approach. Electronic copy available at: http://ssrn.com/abstract=2276927.

[8] Okenwe, I. Opara, J., Ononogbu A. C. and Bassey, U. (2016). Parametric Versus Non-Parametric Simple Linear Regression on Data with and Without Outliers. International Journal of Innovation in Science and Mathematics. Volume 4, Issue 5, Sept. 2016.

[9] Opara, J., Iheagwara, A.I., and Okenwe, I. (2016). Comparison of parametric and non-parametric linear regression. Advance Research Journal of Multi-Disciplinary Discoveries. Vol.2.0/Issue-I

[10] Ranjit, K.P. (2005). Some Methods of Detection of Outliers in Linear Regression Model. Ebook_2005_2006_MSc._trim1_4. Unpublished

[11] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461-464

[12] Sen, P.K., (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association 63 (324): 1379-1389.

[13] Sprent, P. (1993). Applied Nonparametric Statistical Methods. London; New York: Chapman and Hall.

[14] Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis. Nederlandse Akademie Weten chappen Series A 53: 386-392.

[15] Wilcox, R. (1998). Simulations on the Theil-Sen regression estimator with right-censored data. Stat. & Prob. Letters 39, 43-47.

★★★